DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA

TECHNICAL FINAL REPORT

INVESTIGATIONS ON FLUCTUATIONS OF SUMS OF RANDOM VARIABLES.

E. Sparre Andersen

Mathematical Institute
Aarhus University
Aarhus, Denmark

TECHNICAL FINAL REPORT

INVESTIGATIONS ON FLUCTUATIONS OF SUMS OF RANDOM VARIABLES.

E. Sparre Andersen

Mathematical Institute
Aarhus University
Aarhus, Denmark

Investigations on Fluctuations of Sums of Random variables.

Abstract.

We have investigated the fluctuations of sums of random
variables X_1, X_2, ... We have generalized previous results
on the random variables H_n connected with this sequence, ob-
tained new results on the validity of the Arc-sine Law for
independent not identically distributed random variables, ob-
tained a generalization of Spitzer's identity, and obtained a
generalization of the equivalence principle. Furthermore Toep-
litz matrices of Laurent polynomials have been studied. For the
growth of the maximal order statistics of a sequence of inde-
pendent, identically distributed random variables a "Law of
the iterated logarithm" has been obtained.

Investigations on Fluctuations of Sums of Random Variables.

1. The main subject for work under the Contract has been
the investigation of the fluctuations of sums of random vari-
ables. We have worked on different problems within this theory.

2. Let X_1, X_2, ... be a sequence of random variables
and let $S_0 = 0$, $S_k = X_1 + ... + X_k$ for $k = 1, 2, ...$. We
have investigated the behaviour of the largest convex minorant
sequence $Z_{n,k}$, $k = 0, 1, ..., n$ to the sequence S_0, S_1, ...S_n .
The number of equalities $S_k = Z_{n,k}$, $k = 1, ..., n$ is, to some
extent, a measure of the fluctuations of the sequence S_0, ...S_n ,
we denote this number by H_n . (We shall define H_0 as 0).
It is clear that H_n is a random variable, the distribution of
which does depend only on the distribution of $X_1, ..., X_n$.
The meaning of H_n may be clarified by considering figure 1.,
where the points have coordinates (k, S_k), $k = 0, ..., 10$.
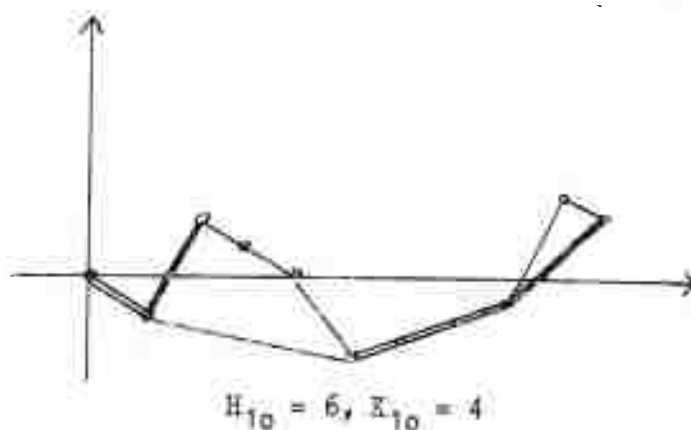

$H_{10} = 6, K_{10} = 4$

Fig. 1.

In the figure H_{10} has the value 6 . It follows from the
figure that another natural statistics similar to H_n is the
number of straight segments in the convex polygon. We denote
this number by K_n , evidently $H_n \geq K_n$. If the random vari-
ables have a continuous distribution, then $P(H_n = K_n) = 1$.

It was shown by Andersen, [3] , that if the random variables X_1, X_2, ... are independent, and have the same continuous distribution, then

(1) $H(s,t) = \sum\limits_{n=0}^{\infty} \sum\limits_{m=0}^{n} P(H_n = m) s^n t^m = (1-s)^{-t}$, $|s| < 1$, $|t| < 1$,

holds for the double-generating function for $P(H_n = m)$. (Actually in [3] the formula looks somewhat different, due to a trivial change in the definition of H_n) .

The main result obtained for H_n under the Contract is: (cf. Technical Note No. 1.)

Theorem 1. Let X_1, X_2, ... be independent, identically distributed random variables. Let x_1, x_2, ... be the set of real numbers x for which $P(S_k = k \cdot x) > 0$ for some $k > 0$, and let

(2) $c_j(s) = \sum\limits_{k=1}^{\infty} P(S_k = k \cdot x_j) s^k$, $j = 1, 2, \ldots$

(3) $c_0(s) = \sum\limits_{k=1}^{\infty} P(S_k \neq k \cdot x_j \text{ for } j = 1, 2, \ldots) s^k$

$$= s(1-s)^{-1} - \sum\limits_{j=1}^{\infty} c_j(s) \ .$$

Then

$$H(s,t) = \sum\limits_{n=0}^{\infty} \sum\limits_{m=0}^{n} P(H_n = m) s^n t^m$$

(4)

$$= \exp\left(t \int_0^s x^{-1} c_0(x) dx\right) \cdot \prod\limits_{j=1}^{\infty} \left(1 - t + t \cdot \exp\left(-\int_0^s x^{-1} c_j(x) dx\right)\right)^{-1} \ .$$

Unfortunately this formula is much too complicated to be useful. Even in the case where the random variables X_1, X_2, ... have only two possible values, say -1 and -1 , each with probability $1/2$, it has not been possible to obtain a simple result. It was shown in TN1 that a simpler formula may be obtained if the common distribution of X_1, X_2, ... have only one discontinuity-point. For this case also the doublegenerating function of $P(K_n = m)$ was found.

3. A fundamental result in the theory of fluctuations of sums of random variables is the so-called Arc-sine Law. Let N_n denote the number of positive sums among $S_1, \ldots, S_n$. We say that the Arc-sine Law holds for the sequence $X_1, X_2, \ldots$ of random variables, if

(5) $\qquad \lim_{n \to \infty} P(N_n/n \leqq x) = \dfrac{2}{\pi} \text{ arc sin } x^{1/2}, \; 0 \leqq x \leqq 1$.

If for an $a \in (0,1)$ we have

(6) $\lim_{n \to \infty} P(N_n/n \leqq x) = \dfrac{\sin \pi a}{\pi} \displaystyle\int_0^x y^{a-1}(1-y)^{-a} dy, \; 0 \leqq x \leqq 1$,

then we say that the generalized Arc-sine Law holds for the random variables $X_1, X_2, \ldots$. For $a = 1/2$ the right hand side of (6) reduces to the right hand side of (5).

Erdös and Kac, [13], proved that the Arc-sine Law holds with $a = 1/2$ if the random variables $X_1, X_2, \ldots$ are independent, have mean 0 and variance 1 and obey the Central Limit Theorem. Andersen proved [3] that (6) holds if the random variables $X_1, X_2, \ldots$ are independent, identically distributed and if

(7) $\qquad\qquad \lim_{n \to \infty} P(S_n > 0) = a$.

It was proved by Spitzer [24] that (7) can be replaced by the weaker condition

(8) $\qquad \lim_{n \to \infty} \dfrac{1}{n} (P(S_1 > 0) + \ldots + P(S_n > 0)) = a$.

It is not known whether Spitzer's result is really stronger than Andersen's result, since it is doubtful that there exist sequences of independent, identically distributed random variables for which the limit in (7) does not exiwt, while (8) is satisfied.

The result of Erdös and Kac apparently connects the
Arc-sine Law with the Central Limit Theorem, whereas the re-
sult of Andersen shows that there is no such connection, as
it is possible to choose for the common distribution of the
random variables X_1, X_2, ... any symmetric, non-degenerate
distribution and thereby obtain the Arc-sine Law.

It is fairly easy to see by heuristic arguments that the
number of positive sums N_n ought to follow the Arc-sine Law,
if the random variables X_1, X_2, ... are in some sense
"asymptotically identically distributed". We have investigated
this problem under the Contract without having obtained a ge-
neral result. We have, however, obtained certain results
which throw some light on the problem.

Theorem 2. Let X_1, X_2, ... be a sequence of independent
random variables. If this sequence is periodic in the sense
that there exist two positive integers p and q and a set
of distribution functions $F_1(x)$, ..., $F_p(x)$, not all degene-
rate, such that $P(X_{np+r} \leq x) = F_r(x)$, $1 \leq r < p$, if $n \geq q$, then
the Arc-sine Law holds for X_1, X_2, ... if $\lim_{n \to \infty} P(S_n > 0) = 1/2$.

Theorem 3. Let X_1, X_2, ... be a sequence of independent,
normally distributed random variables with mean 0. Then

(i)   if $E(S_n^2) = \log n$, $n = 1, 2, ...,$ we have

$$\lim_{n \to \infty} P(N_n/n \leq x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{2} & \text{for } 0 < x < 1 \\ 1 & \text{for } 1 < x \end{cases}$$

(ii) if $E(S_n^2) = n^b \lambda(n)$, $n = 1, 2, ...,$ where $b > 0$ and
$\lambda(x)$ is defined, positive and monotone in $0 < x < \infty$
and for every positive constant c satisfies

$$\lambda(cx)/\lambda(x) \longrightarrow 1 \quad \text{for} \quad x \longrightarrow \infty \quad ,$$

we have

$$\lim_{n \to \infty} P(N_n/n \leq x) = P\left(\int_0^1 \frac{1 + \text{sign } X(t)}{2} \, dt^{1/b} \leq x \right), \quad 0 \leq x \leq 1 \quad ,$$

where $X(t)$, $0 \leq t \leq 1$, is the Wiener Process,

(iii) if $E(S_n^2) = b^n$, $b > 1$, $n = 1, 2, \ldots$, we have

$$\lim_{n \to \infty} P(N_n/n \leq x) = \begin{cases} 0 & \text{for } x < \frac{1}{2} \\ 1 & \text{for } \frac{1}{2} < x \end{cases}$$

A proof of Theorem 2 is given in TN2. Theorem 3 may be proved by a modification of the invariance principle used by Erdös and Kac in $\boxed{13}$. Case (i) of Theorem 3 is due to G. Maruyama $\boxed{22}$, together with (iii) it represents extreme cases where the variance of S_n increases so slowly that practically all sums have the same sign, or so fast that for large n approximately 50 % of the sums are positive. Case (iii) for $b = 1$ gives the usual Arc-sine Law, for other values of b it has not been possible to evaluate

$$P\left(\int_0^1 \frac{1 \pm \text{sign } X(t)}{2} \, dt^{1/b} \leq x \right) .$$

The most interesting case of Theorem 3, therefore, is case (ii) with $b = 1$. It shows that the Arc-sine Law may hold for a sequence of independent random variables $X_1, X_2, \ldots$, even if the variance of X_n goes to infinity like $(\log n)^c$, $c > 0$ for $n \longrightarrow \infty$. In this case the random variables $X_1, X_2, \ldots$ cannot be said to be "asymptotically identically distributed" in the usual sense of the words. From this, and also from Theorem 2, it follows that the Arc-sine Law must hold for a sequence of independent random variables under rather weak assumptions about the distribution functions. Several unsuccessful attempts have been made to prove that the Arc-sine Law holds under conditions which are more general than the conditions in Theorem 2, or case (ii) of Theorem 3. Our investigations have led us to the conjecture that the generalized Arc-sine Law holds for the independent random variables $X_1, X_2, \ldots$ with distribution functions $F_1(x), F_2(x), \ldots$, if

$$\lim_{n \to \infty} P(S_n > 0) = a$$

and if there exists a sequence of distribution functions $G_1(x)$, $G_2(x)$, ... such that to any $\xi > 0$ we can find an $N(\xi)$, tending to infinity when $\xi \longrightarrow 0$, and a $\delta(\xi) > 0$, tending to 0 when $\xi \longrightarrow 0$ such that

$$\sup_{-\infty < x < \infty} |\ P(S_{k+n} - S_k \leq x) - G_n(x)|\ < \xi$$

whenever $k \geq 0$, $k + n \leq N(\xi)$ and $n \geq \delta(\xi)N(\xi)$. The last condition is the way in which we think it natural in this connection to express that the variables X_1, X_2, ... are "asymptotically identically distributed". We have not been able to verify this conjecture.

    <u>4</u>. A well-known result in the theory of fluctuations of sums of random variables is Spitzer's identity, [24],

$$(9) \quad \sum_{n=0}^{\infty} \varphi(T_{n,o}, S_n)s^n = \exp \sum_{n=1}^{\infty} \frac{1}{n}\varphi(S_n^+, S_n)s^n \ , \quad |s| < 1 \ ,$$

where $\varphi(X,Y)$ denotes the joint characteristic function of the random variables X and Y, while $T_{n,o} = \max(S_o, S_1, \ldots, S_n)$, and $S_n^+ = \max(0, S_n)$. Spitzer's identity is valid if the random variables X_1, X_2, ... are independent and identically distributed. Several proofs are known, some combinatorial and others relying mostly on analysis or functional analysis.

    We have treated a generalisation of Spitzer's formula to symmetrically dependent random variables. The random variables X_1, ..., X_n are said to be symmetrically dependent, if the joint distribution function $F_n(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$ is a symmetric function of $x_1, \ldots, x_n$. The random variables X_1, X_2, ... are said to be symmetrically dependent if, for each positive n, the random variables X_1, ..., X_n are symmetrically dependent. We remark that if the random variables X_1, X_2, ... are symmetrically dependent, then they are equivalent according to the terminology used by de Finetti, [16].

It was shown by de Finetti that it is possible to deduce re-
sults for equivalent random variables from results for indepen-
dent random variables. We do, however, prefer to work directly
on symmetrically dependent random variables for two reasons.
One is that the random variables $X_1, \ldots, X_n$ may be symmetri-
cally dependent although it is not possible to extend the finite
sequence $X_1, \ldots, X_n$ to an infinite sequence $X_1, \ldots, X_n, X_{n+1}, \ldots$
such that the infinite sequence of random variables are symme-
trically dependent. We therefore in some cases get more general
results when we work directly on symmetrically dependent random
variables. The other is that the combinatorial methods, we use,
seem to be naturally adopted to the concept of symmetrically de-
pendent random variables.

For independent random variables multiplication of characte-
ristic functions corresponds to addition of random variables.
The same is not true for dependent random variables. If, however,
a set of symmetrically dependent random variables is given, then
it is possible to introduce a symbolic multiplication of cha-
racteristic functions of random variables which are functions
of the given set of symmetrically dependent random variables.
This symbolic multiplication coincides with the ordinary multi-
plication, in case that the given symmetrically dependent ran-
dom variables are independent and identically distributed.
Having introduced this symbolic multiplication of characteristic
functions the definition of the right-hand-side of (9) follows
by application of the exponential series. We have proved by
combinatorial methods that (9) holds for symmetrically depen-
dent random variables $X_1, X_2, \ldots$ when the multiplication of
characteristic functions is the symbolic multiplication intro-
duced above.

$\underline{5}$. The equivalence principle in the theory of fluctuations
of sums of random variables states that the number N_n of posi-
tive sums among $S_1, \ldots, S_n$ has the same distribution as the
index L_n of the first maximum, if the random variables
$X_1, \ldots, X_n$ are symmetrically dependent. We have, in $\lfloor 4 \ \rfloor$,
obtained a generalization of this result. The random variables
$N_{n,k}$ and $L_{n,k}$ are defined by

$N_{n,k}$ = number of indices $j(j=0,\ldots,k-1,k+1,\ldots,n)$

for which $S_j \geqq S_k$ if $j < k$

or $\qquad S_j > S_k$ if $j > k$,

$L_{n,k}$ = index $v(v=0,\ldots,n)$ of the sum S_v

for which exactly k sums

S_j $(j=0,\ldots,v-1,v+1,\ldots,n)$

satisfy $S_j \geqq S_v$ if $j < v$

or $\qquad S_j > S_v$ if $j > v$.

We remark that $N_{n,o}$ is the number of positive sums usually denoted by N_n and that $L_{n,o}$ is the index of the first maximum.

$\underline{\text{Theorem 4}}$. Let X_1, $\ldots$, X_n be symmetrically dependent random variables and let C be an event which is symmetric with respect to X_1, $\ldots$, X_n , then

(10) $\qquad P(N_{n,k}=j,C) = P(L_{n,k}=j,C) \qquad \begin{array}{l} k = 0,\ 1,\ \ldots,\ n \\[4pt] j = 0,\ 1,\ \ldots,\ n \ . \end{array}$

The proof we have given of Theorem 4 in $[\ 4\ ]$ is based on a one-to-one measure-preserving of the n-dimensional sample space onto itself such that the event $[\ N_{n,k}=j,\ N_{k,k}=v,C]$ is mapped onto the event $[\ L_{n,k}=j,\ L_{j,j}=v,C]$. For each point in the sample space the mapping is a permutation of the coordinates.

$\underline{6}$. In connection with the work on fluctuations of sums of random variables we have investigated the Toeplitz matrices of Laurent polynomials. The result obtained are given in TN4. It was conjectured that this study would be valuable for the work on extension of the Arc-sine Law to the case where the random variables X_1, X_2, $\ldots$ form a stationary Markov chain. We have, however, not been able to obtain results in this direction.

7. For a sequence of independent, identically distributed random variables the growth of the maximum of the first n variables have been investigated and a "Law of the iterated logarithm" has been proved in TN5.

Theorem 5. Let X_1, X_2, ... be independent, identically distributed random variables, with distribution function $F(x)$. Let λ_1, λ_2, ... be a nondecreasing sequence of real numbers then

$$P(\lim \sup [\max(X_1,\ldots,X_n) \leq \lambda_n]) = 0$$

if

$$\sum_{n=3}^{\infty} (F(\lambda_n))^n \frac{\log\log n}{n} < \infty$$

furthermore

$$P(\lim \sup [\max(X_1,\ldots,X_n) \leq \lambda_n]) = 1$$

if the sequence $(F(\lambda_n))^n$, $n = 1, 2, \ldots$ is non-increasing and

$$\sum_{n=3}^{\infty} (F(\lambda_n))^n \frac{\log\log n}{n} = \infty.$$

The proof of Theorem 5 is based on a generalization, obtained in TN5, of the convergence part of the Borel-Cantelli lemmas.

Bibliography.

1. E. Sparre Andersen, On sums of symmetrically dependent random variables, **Skand.** Aktuarie tidskr. 36 (1953), 123-138.

2. E. Sparre Andersen, On the fluctuations of sums of random variables, Math. Scand. 1 (1953), 263-285, and Math. Scand. 2 (1954), 193-194.

3. E. Sparre Andersen, On the fluctuations of sums of random variables II, Math. Scand. 2 (1954), 195-223.

4. E. Sparre Andersen, Combinatorial methods in probability theory, Report from Colloquium in Aarhus August 1962, 13-16.

5. Glen Baxter, Notes for a seminar in stochastic processes, University of Minnesota, 1957.

6. M.T.L. Bizley, Derivation of a new formula for the number of minimal lattice paths from (0,0) to (km, kn) having just t contacts with the line my = nx and having no points above this line; and a proof of Grossman's formula for the number of paths which may touch, but do not rise above this line, Journal of the Institute of Actuaries, 80 (1954), 55-62.

7. Achi Brandt, A generalization of a combinatorial theorem of Sparre Andersen about sums of random variables, Math. Scand. 9 (1961), 352-358.

8. W. Doeblin et P. Lévy, Sur les sommes de variables aléatoires indépendantes à dispersion bornées inférieurement, C.R. Acad. Sc. 202 (1936), 2027-2029.

9. M.D. Donsker, An invariance principle for certain probability limit theorems, Mem. Amer. Math. Soc. 6 (1951).

10. A. Dvoretzky, On the strong stability of a sequence of events, Ann. Math. Stat. 20 (1949), 296-299.

11. A. Dvoretzky and Th. Motzkin, A problem of arrangements, Duke Math. J. 14 (1947), 305-313.

12. P. Erdös, On the law of the iterated logarithm, Ann. Math. 2. series, 43 (1942), 419-436.

13. P. Erdös and M. Kac, On the number of positive sums of independent random variables, Bull. Amer. Math. Soc. 53 (1947), 1011-1020.

14. W. Feller, The law of the iterated logarithm for identically distributed random variables, Ann. Math. 2. series, 47 (1946) 631-638.

15. W. Feller, On combinatorial methods in fluctuation theory, The Harald Cramer Volume, Ed. Ulf Grenander, 1959, 75-91.

16. B. de Finetti, La prévision: ses lois logiques, ses sources subjectives, Ann. Inst. H. Poincaré **7 (1937), 1-68.**

17. E. Franckx, Sur la convergence forte, Mitt. Verein. Schweiz. Versich.-Math. 59 (1959), 293-296.

18. J. Geffroy, Contribution à la théorie des valeurs extrêmes (Première partie), Publ. Inst. Stat. Univ. Paris 7 (1958), 37-121.

19. Ulf Grenander and Gabor Szegö, Toeplitz forms and their applications, University of California Press, Berkeley, 1958.

20. M. Kac, On some connections between probability theory and differential and integral equations, Proc. Sec. Berk. Symp. (1951), 189-215.

21. M. Loève, Probability theory, 2nd edition, Van Nostrand, New York, 1960.

22. G. Maruyama, Note on the arc-sine law in the theory of probability, Natural Science Report of the Ochanomizu University 2 (1951), 25-27.

23. Palle Schmidt and Frank Spitzer, The Toeplitz matrices of an arbitrary Laurent polynomial, Math. Scand. 8 (1960), 15-38.

24. F. Spitzer, A combinatorial lemma and its application to probability theory, Trans. Amer. Math. Soc. 82 (1956), 323-339.

25. J.G. Wendel, Spitzer's formula: A short proof, Proc. Amer. Math. Soc. 9 (1958), 905-908.

26. Harold Widom, On the eigenvalues of certain Hermitean operators, Trans. Am. Math. Soc. 88 (1958), 491-522.

Technical Notes and Reports under the Contract.

TN1: E. Sparre Andersen, On the distribution of the random variable H_n (1959).

TN2: O. Barndorff-Nielsen, On certain convergence properties for sequences of indicators (1959).

TN3: cancelled.

TN4: Palle Schmidt, The determinants of the Toeplitz matrices of an arbitrary Laurent polynomial (1960).

TN5: O. Barndorff-Nielsen, On the rate of growth of the partial maxima of a sequence of independent, identically distributed random variables (1961).

TN6: O. Barndorff-Nielsen, Characteristic subsequences and limit laws for weighted means (1962).

TSR: E. Sparre Andersen, On the distribution of the random variable H_n (1959).

TSR2: E. Sparre Andersen, Investigations on the validity of the arc-sine law (1960).

ASR3: E. Sparre Andersen, Investigations on fluctuations of sums of random variables. (1961).

MATHEMATICAL INST
AARHUS, DENMARK
15 July 1963

AF 61(052)-42
TFR
MATHEMATICS

INVESTIGATIONS ON FLUCTUATIONS OF SUMS OF RANDOM VARIABLES

E. SPARRE ANDERSEN

ABSTRACT: Distribution of the r.v. H. Validity of the Arc-sine Law for not identically distributed r.v.'s. Generalization of Spitzer's formula and of the equivalence principle. Toeplitz matrices of Laurent polynomials. Law of the iterated logarithm for the maximal order statistics.

UNCLASSIFIED